Cardiovascular Risk Prediction Summary

Team Member's Name, Email and Contribution:

1. **Mujtaba Ali** – mujtwa@gmail.com

- EDA
- Clean-Up
- Feature Engineering

2. **Prateek Sachdeva** – prateeksachdeva13@gmail.com

- Pre-Processing & dataset balancing
- Model Implementation
- Model Explainability

Please paste the GitHub Repository & Google Drive link.

Github Link:- https://github.com/PrateekSachdevaa/CAPSTONE-PROJECT-3-CLASSIFICATION

Google Drive Link:-
https://drive.google.com/drive/folders/1mV7Xq6V3U8UN2B8OoALCc9p98sE7GeYw

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

- Cardiovascular disease (CVD) is a class of diseases that involve the heart or blood vessels. CVD includes coronary artery diseases (CAD) such as angina and myocardial infarction (commonly known as a heart attack). Other CVDs include stroke, heart failure, hypertensive heart disease, rheumatic heart disease, cardiomyopathy, abnormal heart rhythms, congenital heart disease, valvular heart disease, carditis, aortic aneurysms, peripheral artery disease, thromboembolic disease, and venous thrombosis.
- The underlying mechanisms vary depending on the disease. It is estimated that dietary risk factors are associated with 53% of CVD deaths. Coronary artery disease, stroke, and peripheral artery disease involve atherosclerosis. This may be caused by high blood pressure, smoking, diabetes mellitus, lack of exercise, obesity, high blood cholesterol, poor diet, excessive alcohol consumption, and poor sleep, among other things. High blood pressure is estimated to account for approximately 13% of CVD deaths, while tobacco accounts for 9%, diabetes 6%, lack of exercise 6%, and obesity 5%.Rheumatic heart disease may follow untreated strep throat.
- The dataset is from an ongoing cadiovascular study on residents of the town of Framingham, Masschusetts. The classification goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD). The dataset provides the patients information. It includes over 4000 records and 15 attributes. Each attribute is a potential risk factor. There are both demographic, behavioral, and medical risk factors.

- Starting with loading the data so far we have done EDA, null values treatment, encoding of categorical columns, feature selection and then model building.
- In all of these models our accuracy revolves in the range of 59 to 89%.
- Total cholesterol was found to be the most important feature, which is contributing the highest in predicting the target variable.
- We found 67% accuracy through logistic regression.
- Moreover found 84% accuracy through knn-classifier.
- In support vector machine, we got 78% accuracy.
- With the help of Naïve Bayes, we got 59% accuracy.
- In decision tree classifier, we found 83% accuracy.
- The ensembles of decision tree i.e. Random forest gave us the highest accuracy i.e. 89% in predicting the target variable.
- However Xgboost classifier gave us 85% accuracy.
- So we found Random forest to be the best performing algorithm with an accuracy of 89%.